Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology 41-1

A method of determining the winsorization threshold, with an application to domain estimation

by Cyril Favre Martinoz, David Haziza
and Jean-François Beaumont

Release date: June 29, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 57-77
Statistics Canada, Catalogue No. 12-001-X

A method of determining the winsorization threshold, with an application to domain estimation

Cyril Favre Martinoz, David Haziza and Jean-François Beaumont[1]

Abstract

In business surveys, it is not unusual to collect economic variables for which the distribution is highly skewed. In this context, winsorization is often used to treat the problem of influential values. This technique requires the determination of a constant that corresponds to the threshold above which large values are reduced. In this paper, we consider a method of determining the constant which involves minimizing the largest estimated conditional bias in the sample. In the context of domain estimation, we also propose a method of ensuring consistency between the domain-level winsorized estimates and the population-level winsorized estimate. The results of two simulation studies suggest that the proposed methods lead to winsorized estimators that have good bias and relative efficiency properties.

Key Words: Conditional bias; Robust estimation; Winsorized estimator; Influential values.

1 Introduction

In business surveys, it is not unusual to collect economic variables for which the distribution is highly skewed. In this context, we often face the problem of influential values in the selected sample. These values are typically very large, and their presence in the sample tends to make classical estimators very unstable.

It is possible to guard against the impact of influential values at the design stage by selecting with certainty the potentially influential units. For example, in business surveys, it is customary to use a stratified simple random sampling without-replacement design containing one or more take-all strata that are usually composed of large units. Unfortunately, it is seldom possible to completely eliminate the problem of influential values at the design stage. The strata in business surveys are usually formed using a geography variable, a size variable (for example, number of employees) and a classification variable (for example, the North American Industry Classification System (NAICS) code). In a survey that collects dozens of variables of interest, it is not unlikely that some of them will have little or no correlation with the stratification variables, which may result in the presence of influential values. This is the case in particular in Statistics Canada's environmental surveys, such as the Agricultural Water Survey, one of whose objectives is to measure the quantity of water used by Canadian farms for irrigation. It turns out that water consumption in a given year has little correlation with the stratification variables, since consumption depends in part on the weather conditions affecting the sampled farms. Another example is the Industrial Water Survey, one of whose objectives is to measure the quantity of water used. In the case of mining companies, the consumption of water for ore extraction is strongly correlated with the geophysical characteristics of the land, which are not taken into account by the stratification variables.

Another problem that leads to influential values in the sample is the presence of stratum jumpers, which arises when the stratification information collected in the field is different from the information in

1. Cyril Favre Martinoz, Laboratoire de Statistique d'Enquête, CREST/ENSAI & IRMAR, Campus de Ker Lann, 35170 Bruz, France; David Haziza, Département de mathématiques et statistique, Université de Montréal, Montréal, Canada, H3C 3J7 and Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France. E-mail: haziza@dms.umontreal.ca; Jean-François Beaumont, Statistical Research and Innovation Division, Statistics Canada, Ottawa, Canada, K1A 0T6.

the sampling frame. These differences are usually due to errors in the frame (for example, an outdated frame). A stratum jumper is a unit that is not in the stratum that it would have been assigned to if the information in the frame had been accurate. If a unit with a large value is assigned to a take-some stratum, it will have a large value for the variable of interest and possibly a large sampling weight, which will potentially make it very influential. In practice, it is not unusual to have between 5% and 10% stratum jumpers.

Classical estimators (such as the expansion estimator) exhibit (virtually) no bias, but they can be very unstable in the presence of a influential values. Robust estimators are constructed so as to limit the impact of influential values, which leads to estimators that are more stable but potentially biased. The objective is to develop robust estimation procedures whose mean square error is significantly smaller than that of classical estimators when there are influential values in the population but which do not suffer a serious loss of efficiency when there are none. The treatment of influential values usually strikes a trade-off between bias and variance.

Winsorization is a method often used in business surveys to treat influential values. It involves decreasing the value and/or weight of one or more influential units to reduce their impact. Two forms of winsorization are considered: standard winsorization and the winsorization described by Dalén (1987) and Tambay (1988). These methods are described in Section 4. Whichever type is used, winsorization requires the determination of a constant that corresponds to the threshold above which large values are reduced. The choice of this constant is crucial, as a poor choice may lead to winsorized estimators that have a larger mean square error than classical estimators. The problem of choosing the constant has been studied by Kokic and Bell (1994) and Rivest and Hurtubise (1995), among others. In the case of a stratified simple random sampling without-replacement design, these researchers determined the constant that minimizes the estimated mean square error of the winsorized estimators. For repeated surveys, they suggest using historical data collected in previous iterations. Kokic and Bell (1994) determined the optimal value of the constant by setting up a common mean model in each stratum and minimizing the winsorized estimator's mean square error with respect to both the model and the sampling design. Clark (1995) generalized the results obtained by Kokic and Bell (1994) to the case of a ratio estimator and by calculating the mean square error with respect to the model only.

First, we consider a different criterion, which involves finding the constant that minimizes the largest estimated conditional bias in the sample. As we explain in Section 2, the conditional bias associated with a unit is a measure of influence that takes into account the sampling design used. The proposed method has the advantage of being simple to apply in practice. In addition, unlike the methods proposed in the literature, it does not require historical information or a model describing the distribution of the variable of interest in each stratum. Robust estimation based on the conditional bias is presented in Section 3.

In Section 5, we deal with the problem of domain estimation, which is an important problem in practice. We apply a robust method separately in each domain of interest. A population-level estimator can easily be produced by aggregating the robust estimators obtained at the domain level. However, since it is defined as the sum of estimators that are all biased, the aggregate estimator could have a large bias. This point was raised by Rivest and Hidiroglou (2004). We propose a three-step approach: First, apply a robust method separately in each domain of interest to produce initial estimates. Independently, produce an initial robust estimate at the population level. Lastly, using a method similar to calibration (e.g., Deville and Särndal 1992), modify the initial estimates so as to ensure consistency between the robust estimates obtained at the domain level and the robust estimate obtained at the population level. The problem of

consistency for domains has been studied in the context of small area estimation; for example, see You, Rao and Dick (2004) and Datta, Gosh, Steorts and Maple (2011).

We conclude this section with a discussion of the concept of robustness in classical statistics and robustness in finite populations. In classical statistics, we deal with infinite populations, for which we want to estimate the mean, say. In this context, an outlier is a value that was generated under a different model from the one under which the majority of the observations were generated. The presence of outliers in the sample can be attributed to the fact that the population from which the sample is generated is a mixture of distributions or that some observations are subject to measurement errors. In classical statistics, we usually want to conduct inferences about the population of inliers. The aim is therefore to construct estimators that are robust in the sense that they are not seriously affected by the presence of outliers in the sample. In this context, it is desirable to construct robust estimators that have a high breakdown point and/or a bounded influence function. In finite populations, measurement errors are corrected at the verification stage, and it is assumed that there are none left at the estimation stage. The aim is to conduct an inference about the "total" population, which includes both outliers and inliers. In other words, in contrast to classical statistics, we are not just interested in the population of inliers. In this context, estimators that have a high breakdown point and/or a bounded influence function are generally not appropriate because they can lead to large biases. We will give preference to estimators that are robust in the sense that (i) they are more stable than classical estimators in the presence of influential values and almost as efficient as classical estimators in their absence, and (ii) they converge to classical estimators as the sample size and the population size increase. Simulation studies are presented in Section 6. Section 7 concludes with a discussion.

2 Measure of influence: Conditional bias

Consider a finite population of individuals, denoted by U, of size N. We want to estimate the total for the variable of interest y, denoted by $t = \sum_{i \in U} y_i$. From the population we select a sample S, of (expected) size n, using the sampling design $p(S)$. A classical estimator of t is the expansion estimator, also known as the Horvitz-Thompson estimator, $\hat{t} = \sum_{i \in S} d_i y_i$, where $d_i = 1/\pi_i$ is the sampling weight of unit i and π_i denotes its probability of inclusion in the sample. Although the expansion estimator, $\hat{t}$, is design-unbiased for t, it can be highly unstable in the presence of influential values.

To measure the impact (or influence) that a sampled unit has on the expansion estimator, we use the concept of conditional bias of a unit; see Moreno-Rebollo, Muñoz-Reyez and Muñoz-Pichardo (1999), Moreno-Rebollo, Muñoz-Reyez, Jimenez-Gamero and Muñoz-Pichardo (2002) and Beaumont, Haziza and Ruiz-Gazen (2013). Let I_i be the sample selection indicator variable for unit i such that $I_i = 1$ if $i \in S$ and $I_i = 0$, otherwise. The conditional bias of the estimator $\hat{t}$ associated with a sampled unit is defined as

$$B_{1i}^{\mathrm{HT}} = E_p\left(\hat{t} \mid I_i = 1\right) - t = \sum_{j \in U}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) y_j, \tag{2.1}$$

where π_{ij} is the joint probability of inclusion of units i and j in the sample. In general, the conditional bias (2.1) is unknown, since the values of the variable of interest are observed only for the sampled units. In practice, the conditional bias must be estimated. We consider the conditionally unbiased estimator (for example, see Beaumont et al. 2013):

$$\begin{aligned}\hat{B}_{1i}^{\mathrm{HT}} &= \sum_{j\in S}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_j\pi_{ij}}\right)y_j \\ &= (d_i-1)y_i+\sum_{j\in S, j\neq i}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_j\pi_{ij}}\right)y_j.\end{aligned} \tag{2.2}$$

This estimator is conditionally unbiased in the sense that $E_p\left(\hat{B}_{1i}^{\mathrm{HT}}|I_i=1\right)=B_{1i}^{\mathrm{HT}}$. We make the following remarks on the conditional bias and its estimator: (i) The conditional bias (2.1) and its estimator (2.2) depend on the inclusion probabilities π_i and the joint inclusion probabilities π_{ij}. In other words, the conditional bias is a measure that takes the sampling design into account. (ii) If $\pi_i=1$, then $B_{1i}^{\mathrm{HT}}=0$ and, similarly, $\hat{B}_{1i}^{\mathrm{HT}}=0$. That is, when $\pi_i=1$, unit i is selected in all possible samples, and consequently $E_p\left(\hat{t}|I_i=1\right)-t=E_p\left(\hat{t}\right)-t=0$, since $\hat{t}$ is a design-unbiased estimator of t. A unit selected systematically in the sample therefore has no influence and does not contribute to the variance of $\hat{t}$. (iii) The estimated conditional bias (2.2) depends on the second-order inclusion probabilities, π_{ij}. For some designs, these probabilities may be difficult to calculate, in which case approximations will be used. For sampling designs that belong to the class of high-entropy designs (e.g., Berger 1998), a number of approximations of the second-order inclusion probabilities have been proposed in the literature; for example, see Haziza, Mecatti and Rao (2008). An alternative solution is to calculate approximations of the π_{ij} using Monte Carlo methods; see Fattorini (2006) and Thompson and Wu (2008).

For a stratified simple random sampling design, the conditional bias (2.1) associated with sampled unit i in stratum h is given by

$$B_{1i}^{\mathrm{HT}}=\frac{N_h}{N_h-1}\left(\frac{N_h}{n_h}-1\right)(y_i-\bar{y}_{Uh}), \tag{2.3}$$

where n_h denotes the size of the sample selected in stratum $h, \bar{y}_{Uh}=N_h^{-1}\sum_{i\in U_h}y_i$, and U_h denotes the population of units in stratum h of size $N_h, h=1,\ldots,H$. The estimator of the conditional bias (2.2) reduces to

$$\hat{B}_{1i}^{\mathrm{HT}}=\frac{n_h}{n_h-1}\left(\frac{N_h}{n_h}-1\right)(y_i-\bar{y}_{Sh}),$$

where $\bar{y}_{Sh}=n_h^{-1}\sum_{i\in S_h}y_i$ and S_h is the sample in stratum h.

For a Poisson design, the conditional bias of sampled unit i is given by

$$B_i^{\mathrm{HT}}(I_i=1)=(d_i-1)y_i. \tag{2.4}$$

In contrast to the simple random sampling without-replacement design, the conditional bias (2.4) is known for all units in the sample, since it does not depend on finite population parameters.

3 Robust estimation based on the conditional bias

To guard against the undue influence of certain units, it is advisable to construct robust estimators of the total t, that is, estimators that reduce the impact of the most influential units. We consider a class of estimators of the form

$$\hat{t}_R = \hat{t} + \Delta, \tag{3.1}$$

where Δ is a certain random variable. As we will see in Section 4, the winsorized estimators considered can be written in form (3.1). As in Beaumont et al. (2013), we want to determine the value of Δ that minimizes the maximum estimated conditional bias of $\hat{t}_R$ in the sample. Formally, we are seeking the value of Δ that minimizes

$$\max_{i \in S} \left\{ \left| \hat{B}_{1i}^R \right| \right\}, \tag{3.2}$$

where $\hat{B}_{1i}^R$ denotes the estimated conditional bias of $\hat{t}_R$ associated with sampled unit i. This conditional bias is given by

$$\begin{aligned} B_{1i}^R &= E_p\left(\hat{t}_R \,|I_i = 1\right) - t \\ &= B_{1i}^{\mathrm{HT}} + E_p\left(\Delta \,|I_i = 1\right) \end{aligned} \tag{3.3}$$

which is estimated by

$$\hat{B}_{1i}^R = \hat{B}_{1i}^{\mathrm{HT}} + \Delta, \tag{3.4}$$

where $\hat{B}_{1i}^{\mathrm{HT}}$ is a conditionally unbiased estimator of B_{1i}^{HT}. If we note that Δ is a conditionally unbiased estimator of $E_p\left(\Delta \,|I_i = 1\right)$, it follows that the estimator of the conditional bias (3.4) is conditionally unbiased for B_{1i}^R. In other words, we have $E_p\left\{\hat{B}_{1i}^R \,|I_i = 1\right\} = B_{1i}^R$.

Beaumont et al. (2013) showed that the value of Δ that minimizes (3.2) is given by

$$\Delta_{\mathrm{opt}} = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right),$$

where $\hat{B}_{\min} = \min_{i \in S}\left(\hat{B}_{1i}^{\mathrm{HT}}\right)$ and $\hat{B}_{\max} = \max_{i \in S}\left(\hat{B}_{1i}^{\mathrm{HT}}\right)$. Estimator (3.1) then becomes

$$\hat{t}_R = \hat{t} - \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right). \tag{3.5}$$

Beaumont et al. (2013) demonstrated that under certain regularity conditions, the estimator (3.5) is design-consistent; that is, $\hat{t}_R - t = O_p\left(N/\sqrt{n}\right)$.

4 Application to winsorized estimators

Estimator (3.5) can be written in alternative forms, which can make it easier to implement in some cases. We consider the winsorized form. This form has been widely studied in the literature. As mentioned in Section 1, standard winsorization is distinguished from Dalén-Tambay winsorization.

Standard winsorization involves decreasing the value of units that are above a particular threshold, taking their weight into account. Let $\tilde{y}_i$ be the value of variable y for unit i after winsorization. We have

$$\tilde{y}_i = \begin{cases} y_i & \text{if } d_i y_i \leq K \\ \dfrac{K}{d_i} & \text{if } d_i y_i > K \end{cases} \tag{4.1}$$

where $K > 0$ is the winsorization threshold. The standard winsorized estimator of the total t is given by

$$\begin{aligned} \hat{t}_s &= \sum_{i \in S} d_i \tilde{y}_i \\ &= \hat{t} + \Delta(K), \end{aligned} \tag{4.2}$$

where

$$\Delta(K) = -\sum_{i \in S} \max(0, d_i y_i - K).$$

Hence, the estimator (4.2) can be written in the form (3.1). An alternative is to express $\hat{t}_s$ as a weighted sum of the initial values using modified weights:

$$\hat{t}_s = \sum_{i \in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = d_i \frac{\min\left(y_i, \dfrac{K}{d_i}\right)}{y_i}. \tag{4.3}$$

If $\min(y_i, K/d_i) = y_i$ (that is, if unit i is not influential), then $\tilde{d}_i = d_i$. Thus, the weight of a non-influential unit is not modified. In contrast, the modified weight of an influential unit is less than d_i and may even be less than 1. It is worth noting that a unit with a value of $y_i = 0$ presents no particular problems, since its contribution to the estimated total, $\hat{t}_s$, is zero. In this case, an arbitrary value can be assigned to the modified weight $\tilde{d}_i$.

In the case of Dalén-Tambay winsorization, the values of the variable of interest after winsorization are defined by

$$\tilde{y}_i = \begin{cases} y_i & \text{if } d_i y_i \leq K \\ \dfrac{K}{d_i} + \dfrac{1}{d_i}\left(y_i - \dfrac{K}{d_i}\right) & \text{if } d_i y_i > K \end{cases}. \tag{4.4}$$

This leads to the winsorized estimator of the total t_y :

$$\begin{aligned}\hat{t}_{\mathrm{DT}} &= \sum_{i\in S} d_i \tilde{y}_i \\ &= \hat{t} + \Delta(K),\end{aligned} \tag{4.5}$$

where

$$\Delta(K) = -\sum_{i\in S} \frac{(d_i - 1)}{d_i} \max(0, d_i y_i - K).$$

Estimator (4.5) can also be written in the form (3.1). As in the case of $\hat{t}_s$, an alternative is to express $\hat{t}_{\mathrm{DT}}$ as a weighted sum of the initial values using modified weights:

$$\hat{t}_{\mathrm{DT}} = \sum_{i\in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = 1 + (d_i - 1)\frac{\min\left(y_i, \dfrac{K}{d_i}\right)}{y_i}. \tag{4.6}$$

As in the case of the standard winsorized estimator, the weight of a non-influential unit is not modified. Unlike standard winsorization, Dalén-Tambay winsorization guarantees that the modified weights will not be less than 1. Once again, a unit with a value of $y_i = 0$ presents no particular problems, since its contribution to the estimated total, $\hat{t}_{\mathrm{DT}}$, is zero. In this case, an arbitrary value can be assigned to the modified weight $\tilde{d}_i$.

Since the standard and Dalén-Tambay winsorized estimators are of the form (3.1), the optimal constant K_{opt} that minimizes (3.2) is obtained by solving

$$\Delta(K) = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$$

or

$$\sum_{j\in S} a_j \max\left(0, d_j y_j - K\right) = \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2}, \tag{4.7}$$

where $a_j = 1$ in the case of $\hat{t}_s$ and $a_j = \left(d_j - 1\right)/d_j$ in the case of $\hat{t}_{\mathrm{DT}}$. It is shown in the Appendix that a solution to equation (4.7) exists under the following conditions:

1. $\pi_{ij} - \pi_i\pi_j \leq 0$; and
2. $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \geq 0.$

Condition 1 is satisfied for most one-stage designs used in practice, such as stratified simple random sampling and Poisson sampling. Condition 2 implies that $\hat{t}_R$ must be less than or equal to $\hat{t}$, since by construction, a winsorized estimator cannot be greater than the Horvitz-Thompson estimator. It is generally expected that Condition 2 will be satisfied in most skewed populations encountered in business surveys and social surveys. It is also shown in the Appendix that the solution to equation (4.7) is unique if the above conditions are met and if $y_i \geq 0$ for $i \in S$. The Appendix contains a brief description of an algorithm for finding the solution to equation (4.7).

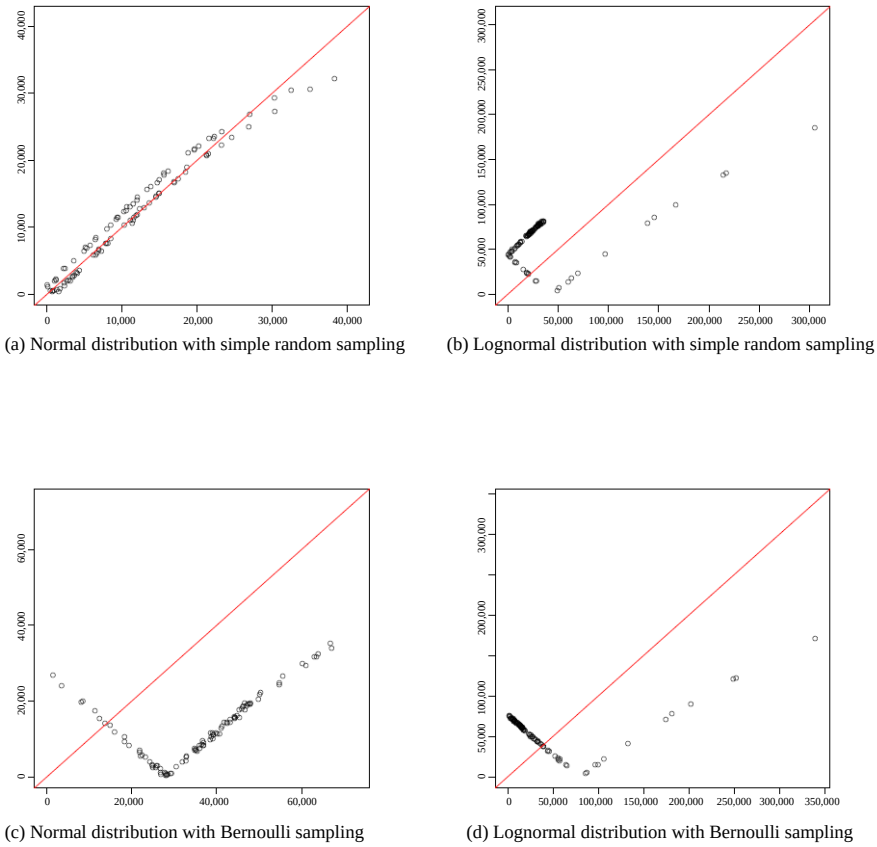
It should be noted that while the value K_{opt} is different for each type of winsorized estimator used, the resulting robust estimators are identical. In other words, we have

$$\hat{t}_s\left(K_{\text{opt}}\right) = \hat{t}_{\text{DT}}\left(K_{\text{opt}}\right) = \hat{t}_R = \hat{t} - \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2}. \tag{4.8}$$

To compare the influence of each population unit with respect to the (non-robust) expansion estimator, $\hat{t}$, and its robust version (4.8), we carried out a simulation study. For that purpose, we generated two populations, each of size $N = 100$. One population was generated according to a normal distribution with mean 4,108 and standard deviation 1,500, and the other was generated according to a lognormal distribution with mean 4,108 and standard deviation 7,373. From each population we selected $M = 500{,}000$ samples according to two sampling designs: (i) a simple random sampling without-replacement design of size $n = 10$, and (ii) a Bernoulli design of expected size $n = 10$. First, we calculated the conditional bias of the Horvitz-Thompson estimator for a simple random sampling without-replacement design, given in (2.3) and for a Bernoulli design, given in (2.4). Note that the conditional bias of the Horvitz-Thompson estimator does not have to be approximated by simulation since all the population parameters are known. The conditional bias associated with unit i of the robust estimator given in (3.3) was approximated as follows: Out of the 500,000 selected samples, we identified those which contained unit i. In each of these samples, we calculated the error, $\hat{t}_R - t$. Finally, we calculated the average value of $\hat{t}_R - t$ over all the samples containing unit i.

The results for the simple random sampling without-replacement design for the normal and lognormal distributions are shown in Figures 4.1 (a) and 4.1 (b) respectively. The results for the Bernoulli sampling design for the normal and lognormal distributions are shown in Figures 4.1 (c) and 4.1 (d) respectively. In each figure, the absolute value of the conditional bias of $\hat{t}_R$ is shown in relation to the absolute value of the conditional bias of $\hat{t}$ for each population unit. The units above the first bisectrix have a conditional bias associated with $\hat{t}_R$ whose absolute value is greater than that of the conditional bias associated with $\hat{t}$. Looking first at the results for simple random sampling without replacement, we see that the behaviour of the absolute value of the conditional bias of $\hat{t}_R$ is similar to that of the absolute value of the conditional bias of $\hat{t}$, which indicates that the influence of the units is not altered significantly after robustification of the expansion estimator. This result is not surprising since the population does not contain any highly influential units. In the case of the lognormal distribution, we see that the influence of the values that have a high conditional bias associated with $\hat{t}$ has been reduced significantly. On the other hand, we note that for the majority of the data, the conditional bias of $\hat{t}_R$ is slightly higher than that of $\hat{t}$. Turning to the results for Bernoulli sampling, we see that in the case of the normal population, the influence of most units has been reduced, since the absolute value of the conditional bias of $\hat{t}_R$ is significantly lower than the

absolute value of the conditional bias of $\hat{t}$. In the case of the lognormal distribution, the results are similar to those obtained with simple random sampling without replacement for the same distribution.



(a) Normal distribution with simple random sampling

(b) Lognormal distribution with simple random sampling

(c) Normal distribution with Bernoulli sampling

(d) Lognormal distribution with Bernoulli sampling

Figure 4.1 Absolute value of the conditional biases of the robust and non-robust estimators

5 Robust estimation of domain totals

In practice, we usually want to produce estimates for population domains as well as an estimate at the global level. Let $t_g = \sum_{i \in U_g} y_i$ be the total of the $y-$variable in domain g. We assume that the domains form a partition of the population such that $t = \sum_{i \in U} y_i = \sum_{g=1}^{G} t_g$, where G is the number of domains. Let S_g be the set of sampled units in domain g. The expansion estimator of t_g is given by $\hat{t}_g = \sum_{i \in S_g} d_i y_i$. We have the consistency relation $\sum_{g=1}^{G} \hat{t}_g = \hat{t}$.

In the presence of influential values, we can apply a robust procedure separately for each domain using the method described in Section 3, which leads to G robust estimators, $\hat{t}_{R,g}$. A robust estimator of the

total at the population level, $\hat{t}_{R(\text{agg})}$, is easily obtained by aggregating the robust estimators $\hat{t}_{R,g}$. Thus, we have $\hat{t}_{R(\text{agg})} = \sum_{g=1}^{G} \hat{t}_{R,g}$. The consistency relation between the domain-level estimates and the population-level estimate is therefore satisfied. However, aggregating G robust estimators, each suffering from a potential bias, may produce a highly biased aggregate robust estimator, $\hat{t}_{R(\text{agg})}$. In most cases, the bias of $\hat{t}_{R(\text{agg})}$ will be negative, since each of the $\hat{t}_{R,g}$ estimators has a negative bias.

To avoid having an estimator with an unacceptable bias, we first compute the robust estimator (4.8), $\hat{t}_{R,g}$, for each domain. Then, we independently compute a robust estimator of the total t in the population, $\hat{t}_{R,0}$, given by (4.8). In this case, however, the consistency relation is no longer necessarily satisfied. In other words, we have $\hat{t}_{R,0} \neq \sum_{g=1}^{G} \hat{t}_{R,g}$, in general. It is therefore necessary to force consistency between the robust domain estimates and the aggregate robust estimate using a method similar to calibration. To do so, we compute final robust estimates $\hat{t}_{R,g}^{*}, g = 0,1,..,G,$ that are as close as possible to the initial robust estimates $\hat{t}_{R,g}$, based on a particular distance function, and that satisfy the calibration equation

$$\sum_{g=1}^{G} \hat{t}_{R,g}^{*} = \hat{t}_{R,0}^{*}. \tag{5.1}$$

In the case of the generalized chi-square distance function, we are seeking final robust estimates, $\hat{t}_{R,g}^{*}$, such that

$$\sum_{g=0}^{G} \frac{\left\{\hat{t}_{R,g}^{*} - \hat{t}_{R,g}\right\}^2}{2q_g \hat{t}_{R,g}} \tag{5.2}$$

is minimized subject to (5.1). The coefficient q_g in the above expression is a weight assigned to the initial estimate in domain $g, \hat{t}_{R,g}$, and is interpreted as its importance in the minimization problem. Using the Lagrange multipliers method, we can easily obtain a solution to this minimization problem. The solution is given by

$$\hat{t}_{R,g}^{*} = \hat{t}_{R,g} - \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \delta_g q_g \hat{t}_{R,g}, \tag{5.3}$$

where $\delta_0 = -1$ and $\delta_g = 1$, for $g = 1,\ldots,G$.

We make the following remarks: (i) If $q_g = 0$, then the final robust estimate $\hat{t}_{R,g}^{*}$ is identical to the initial robust estimate $\hat{t}_{R,g}$. Thus, if we want to ensure that the initial estimate in domain g is not modified excessively, we simply associate it with a small value of q_g. This point is also illustrated empirically in Section 6.2. (ii) Note that like the initial robust estimates at the domain level, $\hat{t}_{R,g}$, for $g = 1,\ldots,G,$ the initial robust estimate at the population level, $\hat{t}_{R,0}$, can also be modified. (iii) If $q_0 = 0$

(in other words, the initial robust estimate for the population level is not modified) and $q_g = q$ for $g = 1, \ldots, G,$ where q is a strictly positive constant, expression (5.3) simplifies to

$$\hat{t}^*_{R,g} = \hat{t}_{R,g}\left(\frac{\hat{t}_{R,0}}{\hat{t}_{R(\text{agg})}}\right). \tag{5.4}$$

In this case, the initial estimates $\hat{t}_{R,g}$ are all modified by the same factor, $\hat{t}_{R,0}/\hat{t}_{R(\text{agg})}$. (iv) How can we set the values of q_g in practice? It seems natural to adopt the following choice:

$$q_g = \widehat{\text{CV}}\left(\hat{t}_g\right)\Big/\sum_{g=1}^{G}\widehat{\text{CV}}\left(\hat{t}_g\right),$$

where $\widehat{\text{CV}}\left(\hat{t}_g\right)$ is the estimated coefficient of variation (CV) associated with domain g. For example, in a repeated survey, the estimated CV observed in a previous iteration can be used. This choice of q_g is based on the fact that we will not want to make a large change in the initial estimate associated with a domain that has a small estimated CV. In such a domain, the problem of influential values is clearly less serious, and the initial robust estimate $\hat{t}_{R,g}$ is expected to be relatively close to the actual total t_g. In other words, the robust estimator $\hat{t}_{R,g}$ should have low bias and be relatively stable. It therefore makes sense not to attempt to change the initial robust estimate substantially. (v) In (5.2), we used the generalized chi-square distance, which leads to the linear method. In the literature on calibration (e.g., Deville and Särndal 1992), there are a number of other calibration methods. In particular, there is the Kullback-Leibler distance, which leads to the exponential method and the logit and truncated linear methods. Using the last two methods, we can specify positive bounds C_1 and C_2 such that $C_1 \leq \hat{t}^*_{R,g}/\hat{t}_{R,g} \leq C_2$. In other words, we ensure that the ratio $\hat{t}^*_{R,g}/\hat{t}_{R,g}$ falls within the interval between C_1 and C_2. Note that the calibration procedure may lead to $\hat{t}^*_{R,g} - \hat{t}_g \geq 0,$ for a certain $g,$ which is counterintuitive. In this case, we simply include the constraint $\hat{t}^*_{R,g} \leq \hat{t}_g$ for $g = 1, \ldots, G,$ in the calibration procedure. (vi) An alternative is to express $\hat{t}^*_{R,g}$ as a weighted sum of the initial values using modified weights:

$$\hat{t}^*_{R,g} = \sum_{i \in S_g} \tilde{d}_i{}^* y_i,$$

where

$$\tilde{d}_i{}^* = \tilde{d}_i\left(1 - \delta_g q_g \frac{\sum_{h=0}^{G}\delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}}\right)$$

and $\tilde{d}_i$ is given by either (4.3) or (4.6). We can also write the estimator $\hat{t}^*_{R,g}$ as a weighted sum with the initial weights using modified values:

$$\hat{t}^*_{R,g} = \sum_{i \in S_g} d_i \tilde{y}^*_i,$$

where

$$\tilde{y}_i^* = \tilde{y}_i \left(1 - \delta_g q_g \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \right), \quad i \in g$$

and $\tilde{y}_i$ is given by either (4.1) or (4.4). (vii) We may want to find the winsorization thresholds $K_g, g = 1, \ldots, G,$ such that the standard winsorized estimator or the Dalén-Tambay winsorized estimator is equal to $\hat{t}_{R,g}^*$. We can follow a procedure similar to the one in Section 4, and we can use an algorithm similar to the one in the Appendix. A necessary condition for the existence of a solution is that $\hat{t}_g - \hat{t}_{R,g}^* \geq 0$. (viii) With the proposed calibration procedure, more than one partition of the population can be dealt with jointly. For example, we may be interested in publishing both provincial estimates and industry estimates. If so, we simply insert the following calibration equations into the calibration procedure:

$$\sum_{g=1}^{G} \hat{t}_{R,g}^* = \hat{t}_{R,0}^*,$$

$$\sum_{l=1}^{L} \hat{t}_{R,l}^* = \hat{t}_{R,0}^*,$$

where G and L denote the number of provinces and the number of industries respectively. The method can also be applied to more than two partitions of the population.

6 Simulation studies

6.1 Winsorization in a simple random sampling without-replacement design

We carried out a simulation study to examine the properties of several robust estimators using 11 populations. The first 10 of size $N = 5{,}000$ consists of a variable of interest y. In each population, the $y-$values were generated according to the following model:

$$Y_i = U_i + \delta_i V_i,$$

where U_i, δ_i and V_i are random variables whose distributions are described in Table 6.1. Population 1 was generated according to a normal distribution. Populations 2 through 5 were generated using a mixture of normal distributions with contamination rates ranging from 0.5% to 5%. Populations 6 through 8 were generated according to skewed distributions. Populations 9 and 10 were generated using a mixture of lognormal distributions with contamination rates equal to 0.5% and 5%. Population 11 of size $N = 5{,}000$ is from the information technology survey produced by the French National Institute for Statistics and Economic Studies (INSEE) in 2011. One of the survey's objectives is to estimate the e-commerce sales of French companies. We use the "sales" variable in our simulation. The distribution of y in each

population is plotted in Figure 6.1. In addition, Table 6.2 presents a number of descriptive statistics for each of the populations used. For confidentiality reasons, the units for Population 11 are not shown in the plot. Similarly, there are no descriptive statistics for Population 11 in Table 6.2.

In each population, we selected $M = 5{,}000$ samples according to a simple random sampling without-replacement design of size $n = 100, 300$ and 500. For each sample, we calculated the expansion estimator $\hat{t}$ and the robust estimator (4.8). Let $y_{(1)}, \ldots, y_{(n)}$ be the values of the $y-$variable arranged in ascending order. We also calculated the first-, second- and third-order winsorized estimators, where the $p^{\text{th}}-$order winsorized estimator is obtained by replacing the p largest values in the sample with the value $y_{(n-p)}$, $p = 1, 2, 3$. In a classical statistical context, Rivest (1994) showed that the first-order winsorized estimator has good mean-square-error properties for a large class of skewed distributions.

As a measure of the bias of an estimator $\hat{\theta}$, we calculated the Monte Carlo relative bias (in percentage):

$$\text{BR}_{\text{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)}{t} \times 100,$$

where $\hat{\theta}_{(m)}$ denotes the estimator $\hat{\theta}$ in sample $m, m = 1, \ldots, 5{,}000$. We also calculated the relative efficiency of the robust estimators with respect to the expansion estimator, $\hat{t}$:

$$\text{RE}_{\text{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)^2}{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{t}_{(m)} - t\right)^2} \times 100.$$

The results are shown in Table 6.3.

The results presented in Table 6.3 show that the once-winsorized estimator has lower bias and is generally more efficient than the two times and three times winsorized estimators, which is consistent with the results obtained by Rivest (1994). It is interesting to compare the robust estimator $\hat{t}_R$ and the once-winsorized estimator. In the case of Population 1, which does not contain any influential values, we see that both estimators have low bias and are as efficient as the expansion estimator. In the case of the populations with a mixture of normal distributions (Populations 2 to 5), we observe that the once-winsorized estimator is less efficient than the robust estimator in every scenario except for Population 5 with $n = 300$. In fact, the once-winsorized estimator is less efficient than the expansion estimator in every scenario except for Population 2 with $n = 100$. The robust estimator is more efficient than the expansion estimator except in Populations 4 and 5, for which we observe values of relative efficiency ranging from 91% to 102%. In the case of the populations with a mixture of lognormal distributions (Populations 9 and 10), we see that the bias and efficiency performance of the once-winsorized estimator and the robust estimator is very similar in all scenarios. The same is true for the skewed populations (Populations 6 to 8), for which the two estimators produce similar results. In the case of Population 11, the robust estimator has a lower bias than the once-winsorized estimator for $n = 100$, though it is less

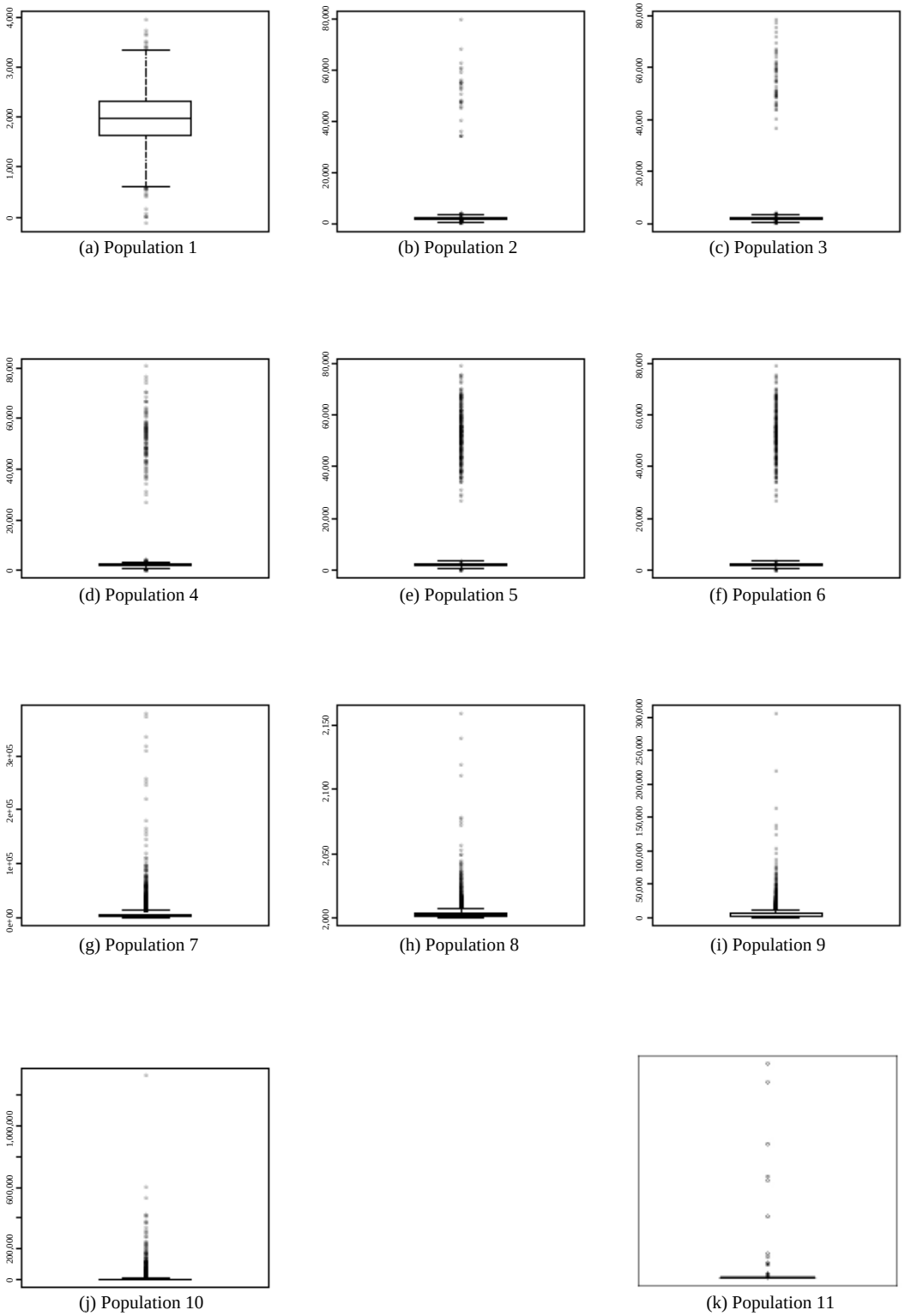
efficient (41% versus 47%). For $n = 300$ and $n = 500$, the robust estimator has a lower bias and is significantly more efficient than the once-winsorized estimator.

Table 6.1
Models used to generate the populations

Population	U_i distribution	Mixture	δ_i distribution	V_i distribution
1	$\mathcal{N}(2{,}000;\ 500)$	No		
2	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.005)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
3	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.01)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
4	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.02)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
5	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
6	$\mathcal{L}\text{og} - \mathcal{N}(\log(2{,}000);1.2)$	No		
7	$\mathcal{L}\text{og}-\mathcal{N}(\log(2{,}000);1.5)$	No		
8	$\mathcal{F}\text{rechet}(2{,}000;2.5;2.1)$	No		
9	$\mathcal{L}\text{og}-\mathcal{N}(\log(2{,}000);1.2)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{L}\text{og}-\mathcal{N}(\log(5{,}000);\ 1.2)$
10	$\mathcal{L}\text{og}-\mathcal{N}(\log(2{,}000);1.2)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{L}\text{og}-\mathcal{N}(\log(5{,}000);\ 1.2)$

Table 6.2
Descriptive statistics for the ten simulated populations

Descriptive	Population									
statistic	1	2	3	4	5	6	7	8	9	10
min	132.3	314.9	105.3	275.9	187.4	23.6	7.6	2,000.9	20.5	26.6
max	3,968	79,506	78,526	80,540	78,690	252,612	379,751	2,159	305,612	1.3×10^{6}
$Q1$	1,639	1,667	1,664	1,666	1,685	883	743	200	920	913
Median	1,986	1,993	1,997	2,015	2,053	1,996	1,981	2,002	2,167	2,041
$Q3$	2,330	2,337	2,339	2,349	2,421	4,505	5,337	2,004	5,018	4,927
Mean	1,985	2,267	2,536	2,976	4,661	4,005	6,118	2,004	4,738	7,883
Standard deviation	503	3,709	5,506	7,119	11,470	7,353	17,190	5.89	9,796	33,111
Skewness	0.0	14.0	10.2	7.3	4.3	4.2	11.6	11.8	12.1	18.4
Kurtosis	3	209	109	56	20	19	196	228	267	570
CV	0.25	1.6	2.2	2.4	2.5	1.8	2.8	2.9×10^{-3}	2.0	4.2



(a) Population 1 (b) Population 2 (c) Population 3

(d) Population 4 (e) Population 5 (f) Population 6

(g) Population 7 (h) Population 8 (i) Population 9

(j) Population 10 (k) Population 11

Figure 6.1 Distribution of the variable of interest in the 11 populations.

Table 6.3
Monte Carlo relative bias (in %) and relative efficiency (in parentheses) of several estimators

Population	n	$\hat{t}_R$	Winsorization		
			Once	Two times	Three times
1	100	-0.1(100)	-0.1(100)	-0.2(101)	-0.3(102)
	300	0.0(100)	-0.0(100)	-0.0(100)	-0.1(100)
	500	0.0(100)	-0.0(100)	-0.0(100)	-0.0(100)
2	100	-4.9(59)	-7.5(87)	-10.7(65)	-11.9(55)
	300	-2.9(87)	-3.0(129)	-6.8(158)	-9.5(169)
	500	-1.9(96)	-1.2(122)	-3.6(175)	-6.5(226)
3	100	-6.9(74)	-8.9(122)	-16.5(119)	-20.0(107)
	300	-3.5(99)	-1.9(122)	-5.6(171)	-10.6(232)
	500	-2.4(102)	-0.9(107)	-2.2(130)	-4.5(186)
4	100	-7.6(91)	-6.2(131)	-15.5(169)	-24.4(194)
	300	-2.9(101)	-0.6(103)	-2.1(118)	-4.4(154)
	500	-2.0(102)	-0.6(102)	-1.1(101)	-1.8(108)
5	100	-5.7(102)	-1.1(104)	-4.1(126)	-9.7(173)
	300	-2.2(102)	-0.4(100)	-0.8(101)	-1.4(102)
	500	-1.2(100)	-0.1(100)	-0.3(100)	-0.5(101)
6	100	-5.7(79)	-5.4(75)	-8.2(80)	-10.6(89)
	300	-2.6(84)	-2.6(79)	-3.9(81)	-5.1(88)
	500	-2.0(86)	-2.0(81)	-3.0(82)	-3.8(88)
7	100	-8.4(72)	-9.3(73)	-14.7(72)	-18.7(79)
	300	-4.5(86)	-4.4(95)	-7.8(91)	-10.2(95)
	500	-3.5(94)	-3.1(105)	-6.0(106)	-8.1(109)
8	100	-0.0(69)	-0.0(75)	-0.0(77)	-0.0(85)
	300	-0.0(82)	-0.0(88)	-0.0(87)	-0.0(95)
	500	-0.0(88)	-0.0(96)	-0.0(94)	-0.0(100)
9	100	-5.7(73)	-5.8(71)	-9.5(72)	-12.4(80)
	300	-3.5(87)	-3.5(85)	-5.4(88)	-6.8(98)
	500	-2.4(88)	-2.4(88)	-3.8(90)	-4.9(97)
10	100	-13.5(68)	-15.0(70)	-24.6(76)	-31.7(89)
	300	-7.5(80)	-7.2(79)	-12.1(85)	-16.3(97)
	500	-5.3(85)	-5.1(83)	-8.4(91)	-11.4(103)
11	100	-22.8(47)	-32.6(41)	-42.0(42)	-47.7(47)
	300	-14.7(65)	-20.0(77)	-29.6(68)	-34.3(75)
	500	-11.3(76)	-14.6(96)	-24.3(90)	-29.3(97)

6.2 Winsorization in a stratified simple random sampling without-replacement design

We also tested the calibration method described in Section 5. We generated a population of size $N = 5{,}000$, which we divided into five strata, $U_1, \ldots, U_5$, of size $N_1, \ldots, N_5$, respectively; see Table 6.4 for the values of N_h. In each stratum, we generated a variable of interest y according to a lognormal distribution with parameters $\log(2{,}000)$ and 1.5.

From the population we selected $M = 5{,}000$ samples according to a stratified simple random sampling without-replacement design. In stratum U_h, we selected a sample S_h of size n_h according to a simple random sampling without-replacement design; see Table 6.4 for the sizes n_h and the corresponding sampling fractions, $f_h = n_h / N_h$.

The objective here is to estimate the total in the population, $t = \sum_{i \in U} y_i$, and the stratum totals $t_h = \sum_{i \in U_h} y_i$, $h = 1, \ldots, H$. In other words, in our example, the strata correspond to domains of interest. Since the strata form a partition of the population, we have the consistency relation, $t = \sum_{h=1}^{H} t_h$. Similarly, the expansion estimators satisfy the consistency relation $\hat{t} = \sum_{h=1}^{H} \hat{t}_h$, where $\hat{t} = \sum_{i \in S} d_i y_i$ and $\hat{t}_h = \sum_{i \in S_h} d_i y_i$ with $d_i = N_h / n_h$ if $i \in U_h$.

For each sample, we first computed the robust estimator (4.8) in each stratum and aggregated the robust estimates to produce an aggregate robust estimate, $\hat{t}_{R(\text{agg})} = \sum_{h=1}^{H} \hat{t}_{R,h}$. Independently, we computed the robust estimator (4.8), denoted $\hat{t}_{R,0}$, at the population level. To ensure that the consistency relation (5.1) was satisfied, we performed the calibration procedure described in Section 5 to obtain the final robust estimates $\hat{t}^*_{R,h}, h = 0, \ldots, 5$. We used four systems of coefficients q_h : (1) $q_0 = 0$ and $q_1 = \ldots = q_5 = 1$; (2) $q_0 = 0$ and $q_h = n_h^{-1}(1 - f_h)$, $h = 1, \ldots, 5$; (3) $q_0 = 0$ and $q_h = \text{CV}(\hat{t}_h) = \sqrt{N_h^2 (1 - f_h) n_h^{-1} S_h^2} / t_h$, where $S_h^2 = (N_h - 1)^{-1} \sum_{i \in U_h} (y_i - \bar{y}_{Uh})^2$, $h = 1, \ldots, 5$; (4) $q_0 = 0$ and $q_h = \widehat{\text{CV}}(\hat{t}_h) = \sqrt{N_h^2 (1 - f_h) n_h^{-1} s_h^2} / \hat{t}_h$, where $s_h^2 = (n_h - 1)^{-1} \sum_{i \in S_h} (y_i - \bar{y}_{Sh})^2$, $h = 1, \ldots, 5$. We make the following remarks on the choice of the coefficients q_h : (i) For all four systems, we assigned a weight $q_0 = 0$ to estimate $\hat{t}_{R,0}$, which is equivalent to making no change in the robust estimate at the population level. In other words, we have $\hat{t}^*_{R,0} = \hat{t}_{R,0}$. (ii) The first weighting system assigns an equal weight to all strata regardless of the sample size or sampling fraction. (iii) In the case of the second system, the coefficient q_h is a function of the sample size n_h and the sampling fraction f_h, but it is independent of the intra-stratum variability S_h^2. (iv) In the third and fourth systems, the choice of q_h depends on the actual CV and the estimated CV respectively, for the reasons mentioned in Section 5.

Table 6.4
Characteristics of the strata

Stratum	1	2	3	4	5
N_h	2,000	1,500	1,000	400	100
n_h	20	75	100	80	80
f_h	0.01	0.05	0.1	0.2	0.8

For each robust estimator, we computed the Monte Carlo relative bias (as a percentage) and the relative efficiency (with respect to the expansion estimator); see Section 6.1. The results are presented in Table 6.5.

The results show that the initial robust estimators $\hat{t}_{R,h}$ are biased, as expected. The bias is larger in strata with a small sampling fraction. For example, in Stratum 1, for which $f_1 = 1\%$, the relative bias of $\hat{t}_{1,h}$ is -11.9%, compared with only -1.5% in Stratum 5, for which $f_5 = 80\%$. We also note that the initial robust estimators are all more efficient than the corresponding expansion estimator, with relative

efficiency values ranging from 57% to 97%. The aggregate estimator $\hat{t}_{R(\text{agg})}$ obtained by summing the initial estimators $\hat{t}_{R,h}, h = 1,\ldots,5$ shows a modest bias with a value equal to -5.7% but is more efficient than the population-level expansion estimator $\hat{t}$, with a relative efficiency of 87%.

The population-level winsorized estimator, $\hat{t}_{R,0}$, shows a small bias with a value equal to -2.8% and is significantly more efficient than the expansion estimator, with a relative efficiency of 81%. The final estimators $\hat{t}^*_{R,h}$ obtained using the system of coefficients $q_h = 1$ for $h = 1,\ldots,5$ all have lower bias than the initial estimator $\hat{t}_{R,h}$, except for Stratum 5. This is due to the fact that we force the sum of the final estimates $\hat{t}^*_{R,h}$ to calibrate on a low-bias estimator. On the other hand, the decrease in the bias is accompanied by a slight decrease in efficiency. For example, in Stratum 4, the relative efficiency is 63% for the robust estimator $\hat{t}_{R,4}$ and 66% for the final estimator $\hat{t}^*_{R,4}$. In the case of Stratum 5, the first system of coefficients is clearly unsuitable, since it leads to a change in the estimate for this stratum, like all the other strata, when this stratum has a very high sampling fraction of 80%. In fact, for this system of coefficients, the estimator $\hat{t}^*_{R,5}$ is less efficient than the expansion estimator, with a relative efficiency of 104. The second choice of coefficients q_h, which takes the sampling fraction f_h and the sample size n_h into account, leads to some interesting results. The final robust estimator in Stratum 1, $\hat{t}^*_{R,1}$, has an appreciably lower bias than the initial estimator $\hat{t}_{R,1}$ and the final estimator based on the first system of coefficients, at the cost of a slight loss of efficiency. For Stratum 5, the estimator $\hat{t}^*_{R,5}$ has a low bias (a relative bias of -0.8%) and the same 97% efficiency as the initial estimator $\hat{t}_{R,5}$. The third and fourth q_h weighting systems lead to similar relative bias and relative efficiency results. For Stratum 1, they lead to lower relative biases than the first weighting system, at the cost of a slight loss of efficiency. For Strata 2, 3 and 4, all four systems of coefficients exhibit similar relative bias and relative efficiency. For Stratum 5, the final estimators are virtually unbiased and no less efficient that the expansion estimator.

Table 6.5
Monte Carol relative bias (in %) and relative efficiency (in parentheses) of the robust estimators at the global level and the stratum level

Global estimator		$\hat{t}_{R(\text{agg})}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$
		-5.7(87)	-2.8(81)	-2.8(81)	-2.8(81)	-2.8(81)
			$\hat{t}^*_{R,h}$			
		$\hat{t}_{R,h}$	$q_h = 1$	$q_h = n_h^{-1}(1 - f_h)$	$q_h = \text{CV}(\hat{t}_h)$	$q_h = \widehat{\text{CV}}(\hat{t}_h)$
Stratum	1	-11.9(57)	-9.1(60)	-0.9(67)	-5.7(62)	-6.7(64)
	2	-6.3(74)	-3.4(76)	-3.3(76)	-3.3(76)	-3.1(78)
	3	-6.0(69)	-3.1(70)	-3.8(69)	-3.2(70)	-3.2(70)
	4	-6.6(63)	-3.7(66)	-4.2(65)	-3.3(66)	-3.4(70)
	5	-1.5(97)	1.5(104)	-0.8(97)	-0.2(98)	0.1(99)

7 Discussion

This paper outlined a proposed method for determining the threshold for winsorized estimators. This method has the advantage of being simple to apply in practice and can be used for sampling designs with

unequal probabilities. We also proposed a calibration method that satisfies a consistency relation between the domain-level winsorized estimates and a population-level winsorized estimate. Although we applied the method in the case of winsorized estimators, it can be used with any type of robust estimator.

Acknowledgements

The authors are grateful to an associate editor and two reviewers for their comments and suggestions, which substantially improved the quality of this paper. David Haziza's research was funded by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

We want to show that there exists a solution to the equation

$$-\Delta(K) = \sum_{j \in S} a_j \max\left(0, d_j y_j - K\right) = \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2} = \hat{t} - \hat{t}_R$$

under the conditions $\pi_{ij} - \pi_i \pi_j \leq 0$ and $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \geq 0$.

First, we arrange the units in order from the smallest value of $b_i = d_i y_i, i \in S,$ to the largest, so that unit 1 has the smallest value of b_i and unit n the largest value. We begin by considering the case of $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) = 0$. We have to solve the equation $-\Delta(K) = 0,$ and we can easily see that this equation is satisfied for all $K \geq b_n$.

We now turn to the case of $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) > 0$. We note first that the function $-\Delta(K)$ is continuous and piecewise linear for $0 \leq K \leq b_n$. The pieces are defined by the intervals $\left[b_{j-1}, b_j\right[, j = 1, \ldots, n,$ where $b_0 = 0$. We also note that $-\Delta(0) = \sum_{j=m}^{n} a_j b_j > 0,$ where m is the smallest index such that $b_m \geq 0$. By the intermediate value theorem, there is a solution to equation (4.7) if we can show that

$$-\Delta(b_n) = 0 < \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j. \tag{A.1}$$

The first inequality follows directly from the condition $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) > 0$. To prove the second inequality, we first note that $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \leq \hat{B}_{\max}$. If we use the estimator of the conditional bias (2.2) and the condition $\pi_{ij} - \pi_i \pi_j \leq 0,$ we observe that $\hat{B}_{\max} \leq (d_k - 1) y_k,$ index k being associated with the unit that has the largest estimated conditional bias. For the Dalén-Tambay winsorized estimator, the last inequality can be rewritten as $\hat{B}_{\max} \leq a_k b_k$. It follows that $a_k b_k \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j,$ which completes the proof that there is a solution to equation (4.7). For the standard winsorized estimator, we can also easily show that $\hat{B}_{\max} \leq a_k b_k$ and therefore that a solution exists. In addition, if the $y_i, i \in S,$ are all positive, the function $-\Delta(K)$ is monotonically decreasing for $0 \leq K \leq b_n$ and the solution is unique.

To find the solution K_{opt}, we find the largest index l such that $-\Delta(b_l) \geq \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$, for $l \leq n$. The solution can then be calculated by linear interpolation between points b_l and b_{l+1}; that is,

$$K_{\text{opt}} = b_l \frac{\Delta(b_{l+1}) - \Delta\left(K_{\text{opt}}\right)}{\Delta(b_{l+1}) - \Delta(b_l)} + b_{l+1} \frac{\Delta\left(K_{\text{opt}}\right) - \Delta(b_l)}{\Delta(b_{l+1}) - \Delta(b_l)},$$

where $\Delta\left(K_{\text{opt}}\right) = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$.

References

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika,* 100, 555-569.

Berger, Y.G. (1998). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference,* 74, 149-168.

Clark, R.G. (1995). Winsorization methods in sample surveys. Masters Thesis, Department of Statistics, Australian National University.

Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R and D Report. Statistics Sweden.

Datta, G.S., Gosh, M., Steorts, R. and Maple, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test*, 20, 574-588.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87, 376-382.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93, 269-278.

Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron,* 66, 91-108.

Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics,* 10, 419-435.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: Conditional bias. *Biometrika,* 86, 923-928.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: Estimating the conditional bias. *Metrika*, 55, 209-214.

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika,* 81, 373-383.

Rivest, L.-P., and Hidiroglou, M. (2004). Outlier treatment for disaggregated estimates. *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, Virginia, 4248-4256.

Rivest, L.-P., and Hurtubise, D. (1995). On Searls' Winsorized mean for skewed populations. *Survey Methodology*, 21, 2, 107-116.

Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, Virginie, 229-234.

Thompson, M.E., and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology*, 34, 1, 3-10.

You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.